Video-Based Assessment of Operative Competency in Laparoscopic Cholecystectomy: A Prospective Observational Study to Inform Rubric Design for AI-Assisted Evaluation

Authors: Yuchen Luo, Chiu Kang, Ankur Sidhu, Mani Suleiman, Nezor Houli, Shekhar Kumta, Russell Hodgson, Northern Health

Northern Health

RESEARCH WEEK
20-24 OCTOBER 2025
INSPIRED RESEARCHERS
Northern Health

◇ Introduction

- Achieving competency in common operations is a key objective of general surgery training, but objectively assessing surgical competency is challenging due to complex technical and non-technical factors and difficulty in standardizing operative performance vision. Existing scoring tools often rely on qualitative assessments and face criticism regarding reliability and validity for high-stakes evaluations.
- This study emphasizes using binary, countable assessment items that cover a range of technical and non-technical factors, with the potential for AI-assisted automated assessment rubric scoring.

Table 3 – Inter-rater agreement comparing numeric variables ICC between mild versus moderate cases.

Items	Mild cases	Moderate/severe cases
Global items		
Hartmann's repetitive grasping	0.64	0.48
Gallbladder repetitive rotation	0.59	0.41
Inefficient instrument insertion	0.68	0.20
Instrument clash	0.33	0.17
Error-based items		
Non-targeted diathermy burning	0.06	0.09
Uncontrolled tissue tearing	0.30	0.12
Dissection into liver	0.08	0.22
Lack of progress over 60 seconds	0.25	0.36
Gallbladder perforation in dissection	1.00	0.81
Gallbladder perforation in excision	0.72	0.46
Procedure-specific items		
Critical view time	0.92	0.93

Note: Strength of agreement reference: No agreement 0; Slight 0 – 0.2; Fair 0.2 – 0.4; Moderate 0.4 – 0.6; Substantial 0.6 – 0.8; Almost perfect 0.8 – 1.0; Perfect 1.0.

Table 1 – Intraclass correlation coefficient ICC analysis for numeric items

Items	Cases	Value	Strength of agreement
Global items			
Hartmann's repetitive grasping	30	0.58	Moderate
Gallbladder repetitive rotation	30	0.53	Moderate
Inefficient instrument insertion	30	0.47	Moderate
Instrument clash	30	0.28	Fair
Non-targeted diathermy burning	30	0.08	Slight
Error-based items			
Uncontrolled tissue tearing	30	0.28	Fair
Incorrectly fired clips	30	0.55	Moderate
Dissection into liver	30	0.17	Slight
Lack of progress over 60 seconds	30	0.31	Fair
Gallbladder perforation in dissection	30	0.89	Almost perfect
Gallbladder perforation in excision	30	0.58	Moderate
Procedure-specific items			
Critical view time	18	0.96	Almost perfect

Note: Strength of agreement reference: No agreement 0; Slight 0 – 0.2; Fair 0.2 – 0.4; Moderate 0.4 – 0.6; Substantial 0.6 – 0.8; Almost perfect 0.8 – 1.0; Perfect 1.0.

Table 2 – Percent agreement for categorical items.

Items	Cases	Percent agreement	Strength of agreement
Procedure-based items			
360-degree dissection	30	0.87	Almost perfect
2 window view	30	0.71	Substantial
Lower third mobilisation	30	0.80	Almost perfect
View of safety	30	0.73	Substantial
Incorrect clipping	30	0.87	Almost perfect
Ability to operate	30	0.64	Substantial
Difficulty level assessment items			
Unrelated adhesion	30	0.98	Almost perfect
Limited view of Hartmann's pouch	30	0.78	Substantial
Gallbladder adhesion	30	0.59	Moderate
Thickened wall	30	0.78	Substantial
Empyema	30	0.87	Almost perfect
Gangrene	30	1.00	Perfect
Intrahepatic anatomy	30	0.93	Almost perfect
Impacted stone	30	0.91	Almost perfect
Unable to grasp	30	0.96	Almost perfect
Distension requiring decompression	30	0.93	Almost perfect
Overall difficulty	30	0.80	Almost perfect

Note: Strength of agreement reference: No agreement 0; Slight 0 – 0.2; Fair 0.2 – 0.4; Moderate 0.4 – 0.6; Substantial 0.6 – 0.8; Almost perfect 0.8 – 1.0; Perfect 1.0.

◇ Methods

Study Design: A prospective cohort study

Participants:

- Four specialist general surgeons (two consultant surgeons, two fellows) and four trainee surgeons.
- Thirty videotapes of laparoscopic cholecystectomy operations.
- Three surgeons invited as raters of these videos

Data Collection & Assessment:

- Assessors received demonstrative training.
- Assessment comprised two parts: delivery of tasks and complexity level of each operative case.
- Case complexity was categorized into mild and moderate/severe difficulty based on features like adhesions, gallbladder appearance, and biliary sepsis, requiring agreement from at least two of the three raters.

Statistical Analysis:

Interrater reliability for numeric items was assessed using the two-way random effects Intraclass Correlation Coefficient model for agreement. Percent agreement was used for categorical items.

◇ Results

Interrater Reliability: The majority of objective measures exhibited at least moderate interrater reliability (ICC, 0.4 – 0.6).

Highest Reliability: Achievement of critical view of safety (PA, 0.71 – 0.87) and case complexity (PA 0.59 – 1.0).

Decline with Complexity: Interrater reliability declined as task complexity increased.

Numerical Variables ICC:

- **Moderately strong agreement (0.47 – 0.58):** Number of times operators grasped Hartmann's pouch (0.58), rotating gallbladder front to back (0.53), incorrect clipping (0.55), redundant movement of instrument insertion (0.47).
- **Almost perfect agreement (0.89 – 0.96):** Gallbladder perforation in dissection (0.89) and time to reach critical view of safety (0.96).

◇ Conclusion

Potential for Objective Tool: There is potential to develop an objective video-based evaluation tool for laparoscopic cholecystectomy based on more quantitative assessment items, supported by the interrater reliability results of this pilot study.

Future Directions: Further research on artificial intelligence and computer vision will be beneficial in establishing a reliable and valid scoring tool. The integration of novel AI technologies holds considerable potential for enhancing the objectivity and precision of surgical training and performance evaluations.